Developing methods for curating multi-omics data

ICAHN SCHOOL OF MEDICINE AT MOUNT SINAI

PI: ZHU, JUN

Grant Number: 1 U01 HG008451-01

Biological systems employ multiple levels of regulation that enable them to respond to genetic, epigenetic, genomic, and environmental perturbations. Advances in high throughput technologies over the past several years have enabled the generation of comprehensive data sets measuring multiple aspects of biological regulation (such as genetics, epigenetics, transcriptomics, metabolomics, glycomics, proteomics, etc.). Many databases, such as TCGA (The Cancer Genome Atlas) database and the LGRC (Lung Genome Research Consortium) database, have been created for depositing diverse types of omics data and for sharing data for public dissemination. However, data errors, including sample swapping, mis-labeling, and improper data entry, during large-scale data generation and data management are inevitable. Our preliminary results indicate that sample labeling errors frequently occur in every database we examined. Data quality control (QC) is critical for all public databases. Data errors need to be identified and corrected before data is released for data analysis and data mining. Analyzing error infested data wastes public resources. Importantly, wrong data could lead to wrong scientific conclusions. And, sample errors could have a large impact on statistic power. To maximally utilize genetic, genomic, and other omics data in public databases, it is critical to properly match different types of data pertaining to the same sample or individual before applying integrative analyses. There is an urgent need for developing methods that can identify data labeling errors in large databases and properly connect diverse types of omics data pertaining to the same individual. In respond to the Big Data to Knowledge (BD2K) initiative, we will develop computational methods to address the topic area ""Data Wrangling"". Here we propose to develop a sample mapping procedure called MODMatcher (Multi- Omics Data matcher) to simultaneously QC multiple types of omics data (Aim 1), and to develop a suite of predictive models based on multi omics data to identify inconsistency between clinical data and omics data (Aim 2). Our proposed methods will be used to clean data, identify and correct data annotation and metadata attribute errors in large databases, which are all within the scope of the ""Data Wangling"". PUBLIC HEALTH RELEVANCE PUBLIC HEALTH RELEVANCE: Sample labeling errors frequently occur in biomedical research databases with diverse types of omics data. We will develop methods to identify and correction data errors in public databases by simultaneously analyzing multiple types of omics data, which are all within the scope of the ""Data Wangling"".